Pre-course Results from the Astronomy Diagnostic Test

Beth Hufnagel1,2,3, Timothy Slater4, Grace Deming1, Jeff Adams4, Rebecca L. Adrian5, Christine Brick6 and Michael Zeilik7

1Department of Astronomy, University of Maryland, College Park, MD 20742-2421, USA
hufnagel@astro.umd.edu
2Department of Physics, University of Maryland, College Park, MD 20742, USA
3Present address: Anne Arundel Community College, 101 College Parkway, Annapolis, MD 21012-1895, USA
4Department of Physics, Montana State University, Bozeman, MT 59717-3840, USA
5Department of Physics and Astronomy, University of Nebraska, Lincoln, NE 68588-0111, USA
6Department of Earth Sciences, Montana State University, Bozeman, MT 59717-3840, USA
7Department of Physics and Astronomy, University of New Mexico, Albuquerque, NM 87131, USA

Received 1999 August 20, accepted 2000 March 24

Abstract: We present selected results from the January 1999 semester pre-course administration of the Astronomy Diagnostic Test (ADT), a research-based, multiple-choice instrument that assesses student knowledge and understanding about selected concepts in astronomy. The ADT is valid for undergraduate non-science majors taking an introductory astronomy course. This paper briefly summarises the development and validation processes, which included pre-course administration to 1557 students in 22 classes attending 17 various post-secondary institutions across the USA in the January 1999 semester. Two interesting results of the ADT’s pre-course administration are (1) the average class score of the ADT is about the same (32%) regardless of type of post-secondary institution or class size and (2) there is a significant gender difference, with women scoring an average of 28% and men 38%, with the standard errors both less than 1%. The current version of the ADT (Version 2 dated 21 June 1999) and a comparative by-class database is available to astronomy instructors at the (USA) Association of Astronomy Educators’ and the National Institute for Science Education’s (NISE) WebPages.

Keywords: astronomy education

1 Why does Astronomy need a Standard Assessment Instrument?

A standard diagnostic test can be a powerful tool to assess the understanding of students, as has been proven for undergraduate physics instruction over the last ten years (e.g. Redish & Steinberg 1999). Several forces are now driving change in undergraduate astronomy education in the USA. These include: a call for post-secondary faculty to document the effectiveness of their teaching; assessment of the relative effectiveness of alternative teaching strategies; and the inclusion of astronomy concepts in the (voluntary) national science education standards for elementary and secondary students (see the analysis by Adams & Slater 1999). For example, a standard baseline should be established before adopting a more interactive teaching style such as that used by Eric Mazur (1997) or Michael Zeilik and his collaborators (1997). These forces are not restricted to the USA, as the UK’s Beyond 2000, Science education for the future, sets forth the expectation that ‘Any contemporary science curriculum . . . will require the development of tools to aid teachers to use formative assessment to monitor and improve pupils’ learning . . .’ (Millar & Osborne 1998). For those simply wishing to teach better, ‘Learning to teach should involve developing the skills of gathering information . . .’ (Hammer 1996).

In July of 1998, a team of astronomy education researchers formed the Collaboration for Astronomy Education Research (CAER) with the goal of producing a multiple-choice, education research-based assessment instrument for introductory, post-secondary astronomy courses for non-science majors. CAER included the authors and a number of others who helped as requested.

2 Development and Validation of the ADT Version 2

The ADT drew from two predecessor surveys. The first was a 47-item multiple-choice instrument developed by Philip Sadler for studying children’s ideas (Sadler 1992) and is referred to as the Project STAR Astronomy Concept Inventory (‘the STAR Inventory’). Although there were publications using it (e.g. Lightman & Sadler 1993; Sadler 1998), it is not widely available. The second predecessor was the Misconceptions Measure, developed by Michael Zeilik and collaborators in 1995 for universities’ large-student-enrollment introductory classes (Zeilik, Schau & Mattern 1998). At the Astronomical Society of the Pacific’s 1998 annual meeting, Zeilik released the Astronomy Diagnostic Test (ADT) Version 1.0. This was a 23-question multiple-choice
instrument, consisting of 13 questions from the previously published Misconception Measure (Zeilik, Schau & Mattern 1998) plus ten additional questions; ten questions in total were from the STAR Inventory. ADT Version 1.0 was the logical starting point for the CAER effort, progressing through Versions 1.1 and 1.9 to produce ADT Version 2.0, a research-based, multiple-choice instrument validated within the USA.

The ADT was re-written by CAER for the September semester of 1998 using standard psychometric principles (e.g. Miyasak & Ryan 1997). These included testing for only one concept per question, avoiding scientific jargon, and allowing the correct answer to be determined without first reading the alternative answers offered. Also, several questions on new topics were added from a previous unpublished survey by one of us (GD), resulting in ADT Version 1.1.

Three types of data dictated improvements to the ADT Version 1.1 to create Version 1.9: statistical analyses from administering Version 1.1, undergraduate interviews, and written responses from students. In the September semester of 1998, the ADT Version 1.1 was administered pre- and post-course to about 1000 students in four colleges and universities, enrolled in eight introductory astronomy classes. The results of statistical analyses, particularly average class scores, item difficulty, and item discrimination, guided rewriting and deletion of questions. We used the answers expressed by students in thirty semi-structured interviews and in thirty written responses as they answered the ADT questions (no answers were provided) to test validity, suggest new questions and new distractors for old questions, identify clearer presentations, and interpret the statistics. All sixty students were enrolled in introductory astronomy at Montana State or the University of Maryland. Several questions from one of us (RLA), based on other interviews with undergraduates (unpublished), replaced uncorrectable questions. The result was ADT Version 1.9 for January 1999.

The basic procedure described above to validate ADT Version 1.1 was repeated in the semester beginning January 1999, with the exceptions of deleting open-ended student responses and adding expert responses. In January 1999, the ADT Version 1.9 was administered pre-course in 17 colleges and universities across the USA to about 1500 students enrolled in 22 introductory astronomy classes, and also taken by their instructors. (Only two classes in one CAER institution were included.) The expert group scored an average of 97%, verifying that there was one correct answer for each question. We used the results from structured interviews of 20 students, as they answered the ADT in multiple-choice format, to determine measurement validity, identify mistakes in typing and image placement (these were minimal), and interpret the statistics. One question was deleted: no new questions were added. The result was ADT Version 2, released on 21 June 1999.

### Table 1. Pre-course ADT scores by institution type

<table>
<thead>
<tr>
<th>Type of Institution</th>
<th>Number of classes</th>
<th>Number of students</th>
<th>Mean score (%)</th>
<th>Standard deviation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>State universities</td>
<td>11</td>
<td>1180</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>Liberal arts</td>
<td>5</td>
<td>237</td>
<td>32</td>
<td>5</td>
</tr>
<tr>
<td>Community colleges</td>
<td>4</td>
<td>95</td>
<td>31</td>
<td>3</td>
</tr>
<tr>
<td>Technical</td>
<td>1</td>
<td>23</td>
<td>51</td>
<td>23</td>
</tr>
<tr>
<td>Women’s</td>
<td>1</td>
<td>22</td>
<td>30</td>
<td>8</td>
</tr>
</tbody>
</table>
Figure 1—Pre-course ADT mean scores from the 22 classes of the January 1999 sample, when placed into 11 bins by class size, are about the same.

Figure 2—Distribution of the binned individual ADT scores for the 683 men is different from the strongly skewed distribution of the 825 women.
The number of students in the 22 classes ranged from six to 201. However, the average pre-course ADT score did not depend on the class size, as Figure 1 illustrates. There is, however, a significant difference in the average ADT scores for men and women. The 683 men in the sample had a mean ADT score of 38%, with a standard error of 0.6%, but the 825 women had a mean score of 28% with a standard error of 0.4%. Figure 2 shows that the binned scores for males and females have different distributions. The symmetrical distribution of the males’ scores shows that their basic understanding has been fairly sampled, but the strongly skewed distribution for the women shows that the ADT is too difficult for them. Moreover, in every one of the 22 classes, the average score for the men was higher than the average score for the women. This confirms a similar pre-course male–female gap found by Zeilik, Schau & Mattern (1999) with students from a university not included in the January 1999 population.

4 Conclusions
The similarity of these pre-course average ADT scores for 22 classes in 17 different institutions is striking. This demonstrates that it is feasible to develop one astronomy diagnostic for non-science majors across the entire USA, regardless of the various elementary and secondary systems which these students experienced, and the post-secondary institution in which they enrolled. Possible exceptions are technical schools, which do not teach a significant number of the non-science majors taking introductory astronomy, and which were represented by only one class in this sample. (In 1996–97 there were 167,800 introductory astronomy course enrollments in US departments with astronomy degree programs—see Mulvey & Nicholson 1999).

The gap between average ADT test scores between men and women is also striking in its size and persistence across types of institution across the USA; one goal of effective astronomy instruction should be to close this gap. Gender differences on individual questions on the ADT suggest that women would disproportionately benefit from additional instruction in selected concepts, e.g. the change in apparent linear size with distance (Hufnagel et al. 2000).

Any standardised survey, particularly a multiple-choice one, has limited usefulness. For example, the ADT should not be used as a graded test, or to assess the abilities of individual students. It cannot reliably assess any one concept, as that would require multiple questions on one concept. It also may not predict student course success for a number of reasons, as discussed by McDermott (1984). The ADT is not intended to guide content selection, nor does it represent a fair sample of typical course content.

A comparative database of the pre-course ADT scores for these 22 classes can help other users of the ADT assess class preparedness for the course as they plan to teach it, and compare his or her class results to other classes in the USA. The ADT and its comparative database are available to astronomy educators at http://solar.physics.montana.edu/aae/adt/ and at the USA’s National Institute for Science Education (NISE) website.

Acknowledgments
This research was supported in part by the National Science Foundation (USA) Grant DGE-9714489 (BH), by NASA Grant CERES-NAG54576 (TS) and by the generosity of the participating astronomy instructors.

References
Adams, J., & Slater, T. 2000, J. Geoscience Education, 48, 39
McDermott, L., 1984, Phys. Today, 37(7), 4
Redish, E. G., & Steinberg, R. N. 1999, Phys. Today, 52(1), 24
Sadler, P. M. 1992, Ph.D Dissertation, Harvard University
Zeilik, M. 1998, personal communication